# Providing a Catalogue of Language Resources for Commercial Users

Maegaard, Bente; Henriksen, Lina; Povlsen, Claus; Olsen, Sussi; Joscelyne, Andrew; Lusicky, Vesna; Mazura, Margaretha; Wacker, Philippe

*Published in:*
Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)

*Publication date:*
2016

# Providing a Catalogue of Language Resources for Commercial Users

**Bente Maegaard[1], Lina Henriksen[2], Andrew Joscelyne[3], Vesna Lusicky[4], Margaretha Mazura[5], Sussi Olsen[6], Claus Povlsen[7], Philippe Wacker[8]**

[1,2,6,7]CLARIN-DK, University of Copenhagen, [3,8]LT Innovate, Brussels, [4]University of Vienna, [5]EMF

E-mail: bmaegaard@hum.ku.dk, linah@hum.ku.dk, aj@lt-innovate.eu, vesna.lusicky@univie.ac.at, mm@emfs.eu, saolsen@hum.ku.dk, cpovlsen@hum.ku.dk, phw@lt-innovate.eu

**Abstract**

Language resources (LR) are indispensable for the development of tools for machine translation (MT) or various kinds of computer-assisted translation (CAT). In particular language corpora, both parallel and monolingual are considered most important for instance for MT, not only SMT but also hybrid MT. The Language Technology Observatory will provide easy access to information about LRs deemed to be useful for MT and other translation tools through its LR Catalogue. In order to determine what aspects of an LR are useful for MT practitioners, a user study was made, providing a guide to the most relevant metadata and the most relevant quality criteria. We have seen that many resources exist which are useful for MT and similar work, but the majority are for (academic) research or educational use only, and as such not available for commercial use. Our work has revealed a list of gaps: coverage gap, awareness gap, quality gap, quantity gap. The paper ends with recommendations for a forward-looking strategy.

**Keywords:** LR, catalogue, users

## 1. Introduction

Language resources (LR) are indispensable for the development of tools for machine translation (MT) or various kinds of computer-assisted translation (CAT). In particular language corpora, both parallel and monolingual are considered most important for instance for MT, not only SMT but also hybrid MT. But corpora are expensive and labour-intensive to create or adapt e.g. for MT usability. Furthermore the extent of availability of LRs differs considerably from language to language. It is true that LRs have been created by EU and national projects and institutions, but they and information about them are scattered across Europe. In order to remedy this lack of overview for the professional user, it is important to apply a user-driven approach towards the identification and mapping of best practices in terms of collecting relevant LT resources.

The Language Technology Observatory will provide easy access to information about LRs deemed to be useful for MT and other translation tools. In order to determine what aspects of an LR are useful for MT practitioners, a user study was made, providing a guide to the most relevant metadata and the most relevant quality criteria. In addition, knowledge and best practice has been extracted from previous studies (LetsMT!, META-SHARE etc.) on collecting relevant LT resources.

## 2. Related work

We have taken the point of departure in existing catalogues and repositories. These comprise: CLARIN VLO (the Virtual Language Observatory, a search facility of metadata for language resources, META-SHARE, ELRA Catalogue of Language Resources, OPUS - the open parallel corpus, TAUS Data, a platform for sharing language data, LetsMT!, LIDER), FALCON 'localization web', PANACEA, a factory of Language Resources (LRs), TTC - Terminology extraction, translation tools, CESAR, EUROTERMBANK, JRC – Joint Research Centre.

## 3. User study

Apart from identifying existing catalogues and existing resources, the consortium has conducted a limited user study in the language resource (LR) user base of EU stakeholders through interviews. A Dialogue Day in Brussels June 2015, Charrette workshops in Vienna (July 2015) and in Brussels (December 2015) contributed to this purpose as well.

The general concerns and facts about the availability, quality and usability of LRs for the commercial sector are shared among the user base at large – *free (in the sense of freely accessible, not necessarily without costs, but with a reasonable cost), good and usable resources* are needed. And obviously the perspective of identifying and making available LRs for commercial & administrative users will have to start with the *user situation, i.e.* to enable potential end users of LRs to access precisely those LRs that fit their purpose. So, from the existing catalogues only relevant resources should be identified.

We have looked into the needs of the following verticals: **Construction** (a major field in Europe, faced with many standards and regulations and much cross-lingual communication due to competition in the sector among major building companies, **Healthcare** (another complex industry with numerous areas requiring translation), **Media monitoring** (for security, marketing, and other business needs), **Procurement** (of interest to the DSI constituency), **Legal and financial** (wide ranging multi-form needs in a competitive business sector). But all relevant domains are taken into account. Dialogue with

users is continuing.

### Input from the users: perspectives on Evaluating LRs

It appears that data on translation needs in verticals is often only available from the language service providers (LSPs) who provide translation services to specific industries. It is very difficult to reach an in-depth understanding of the precise needs of given verticals due to the fact that LSPs are very cautious in releasing this data on which they build their competitive edge.

We have nevertheless been able to formulate provisional conclusions about user needs and the dialogue with users is continuing.

We have investigated the user point of view about the quality and usability of LRs in automated translation contexts by focusing on language service or language technology suppliers who work in this domain. Most of these have a number of clients operating across a broad range of industry domains and text types. The result is impressionistic but effective. This survey work has mainly been carried out through face to face interviews with LR users, some use case research, and consultations with, for example, members of the TAUS community, who systematically use LRs collected in the TAUS repository. Surveys of language industry bodies will also deliver further data on current practices and desiderata with respect to LRs.

There is currently no clear answer from industry about any shared method for evaluating the *quality* of LRs. There are many aspects of the quality of language resources, and often quality and usefulness for a specific task are interrelated.

It is widely thought that LR evaluations using some simple list of key parameters for ratings could be crowd-sourced from the user community. But it will inevitably be time-consuming and partial. This is why in commercial contexts LSPs ask for all the data from their customers and then see what works.

It can be noted that ELRA has a full *Validation* procedure for LRs. What is measured is adherence to the standards used, exhaustiveness etc. The validation process is formal (can be checked automatically) as well as manual. See e.g. http://elra.info/en/services-around-lrs/validation/standards-best-practices/: This does not address the value of a language resource for a specific purpose, but it addresses the soundness of the resource as such. This seems to be as far as quality can be measured.

As in our work we are also relying on existing and acknowledged catalogues, we have assumed that their quality criteria are acceptable also to our users. [1]

[1] For further investigation of the evaluation question we also collaborate with sister projects where applicable. But in the first instance we see that users want to evaluate if a resource is useful for them.

## 4. Defining LR Usability

Usability is a complex but vital criterion for evaluating and using LRs in the context of creating a one-stop shop LR Catalogue service for the business community (one of the objectives of the LT_Observatory project), where relevant information, time and quality are key values for decision-making.

While LRs for research or academic use are fairly abundant (albeit not in all domains or language pairs), LRs for commercial (professional) use are far less easily available. Many LRs available in the repositories surveyed are not available for use other than research. This impedes their commercial use altogether or, when the latter is exceptionally allowed, the pricing conditions are often prohibitive. The next dilemma occurs when trying to evaluate if a LR addresses the domain one is targeting. "Browsing" a sample is hardly ever possible and information about contact persons to obtain further information in a timely manner is often not available and/or not up-to-date. These are very real impediments that limit the operational use of LRs in a commercial context drastically.

The principle aspects of usability on which users are focusing, are:

- ease of access & download: this includes simplicity of access (discoverability, number of clicks, ease of payment, straightforward licensing conditions, etc.), availability of up-to-date contact information
- domain relevance: the metadata must include information on the specific domain covered by the LR and there must be a way of testing (e.g. through samples) to what a LR is domain relevant.
- language pairs: information about language pairs should be available and the depth of coverage by language pair should be clarified.
- availability/cost: cost is an important factor, however, a LR will be considered commercially relevant despite of its cost if all other usability criteria are fulfilled and if the cost is in line with the business models of the language industry
- time to implementation: ultimately, a LR will only be used in a commercial context if the time to implementation does not make its use prohibitive

If these criteria are fulfilled, a LR is considered as "operationally usable".

Usability in this context is understood as set of criteria facilitating human decision making. It is entirely possible that language resources will in the future be selected automatically by digital services operating as part of a broader and deeper language and translation infrastructure, as indicated by some of the EC-funded projects (LIDER and FALCON) listed above. It is to be anticipated that the above criteria will nevertheless stay relevant if this trend was confirmed.

The above described usability criteria based on user needs have been identified in our discussions with stakeholders. They have been taken into account in the methodology chosen for LR collection in view of putting together our LR Catalogue, hosted on LT-Innovate's web platform.

## 5. Metadata discussion

It is essential that the LTO catalogue provides users with easy access to the resources they need. This means that the search options of the LTO catalogue must accommodate exactly the information types (metadata) that users are interested in. Therefore the metadata categories selected for the LTO are based on the user study, the previous usability check list as well as on experiences from similar projects.

Similarly, it is important that the language resources selected from various existing catalogues for inclusion in the LTO catalogue are exactly the resources that users need for MT-purposes. They must therefore be chosen by means of carefully prepared selection criteria.

We have compared the use of metadata in three major projects: LetsMT!, CLARIN VLO and META-SHARE.

## 6. LTO metadata categories

Building on the knowledge described in section 5, combined with the user input, we agreed from the beginning on a minimal list of metadata: Title (of the resource), Type of resource (corpus, terminology, lexicon etc.), Creator, Language(s), Availability (available for commercial use, price) , Modality (written for the time being), URL, Domain, Format (e.g. plain text), Size (in words, or any other measure), Production date, Comment (here additional information can be stored). Experience gained from the collection process and from user feedback made us add the following metadata fields: Description, Tags, Contact person, and Format description. Description and tags turned out to be real added value for many resources, contact person is only available for a part of the resources, and format description is hardly ever filled in.

## 7. Selection criteria for collection of language resources

Based on user input and on experience from previous projects, the below selection criteria have been used as a framework for extraction of useful language resources.
The list below shows the most important selection criteria with a rating from 0-2 where 0 is the least accepted/desired value.

**Availability**
2 – the resource is available and it is freely, openly available under sensible license.
1 – the resource is potentially available but its licenses need negotiation; there may be a cost.
0 – the resource has restricted access.
**Languages covered**
2 – bilingual, multilingual
1 – monolingual
And we need to cover as many languages as possible
**Longevity**
1– resource is actively maintained; the current version is less than 5 years old.
0 – resource is unmaintained.
**Validation** of resource
2 – extensively tested;
1 – moderately tested resource
0 – untested.
**Modality** of resource
1 – text – at present we are only selecting written text
0 – other
**Ease of download**
1 – it is easy to download/obtain the resource, at most 3 clicks
0 – too heavy

## 8. Valorisation process

Some of the collected resources were originally created in contexts where metadata were for various reasons not given high priority. These resources thus lack information types as e.g. domain, language, resource type and/or creator – and consequently potentially valuable language resources are practically more or less unusable.

The valorization of language resources therefore proved crucial and included more steps. All the collected resources have been through a validation process with review of all links and existing metadata. Another important step of the valorization concerns the invitation to user feedback through workshops and through a user response system embedded in the LTO online resource. This way, potential and actual users are asked to give ratings about the usability of LTO language resources and suggest possible improvements.

New insights gained through the validation process and through user feedback steps have resulted in:

- optimization of existing metadata
- addition of new metadata (e.g. resource description, tags and format description)
- extension of language resource selection criteria to also include: high priority to manually validated resources, more focus on high-quality monolingual corpora, preference to TMX and XLIFF data formats.

## 9. Tools for creating corpora

More investigations point to the fact that inclusion of small amounts of in-domain parallel data can improve significantly the translation quality of SMT systems (Pecina et al, 2012, Mastropavlos & Papavassiliou, 2011). Bearing in mind the huge amounts of documents available online, it would be obvious to define and implement methods that identify and acquire domain specific bilingual corpora from the Web. Creating such corpora is relevant in particular when targeting less resource covered languages.

In broad terms, acquisition of in-domain parallel data can be divided into three phases. The first step consists of a

focused search for and subsequently ranking of domain relevant websites. The links found at these websites are then regarded as candidate URL seeds with respect to identifying bilingual documents. After evaluation of the candidate documents detected, the final process consists of removal of duplicates, exclusion of boilerplate elements, extraction of parallel sentences, tokenization, and finally sentence alignment. Several of the elements in this workflow can be treated by existing, open source tools. To give some examples, at http://nlp.ilsp.gr/soaplab2-axis research tools can be downloaded that find links within websites and at https://github.com/danielvarga/hunalign the Hunalign sentence aligner can be downloaded (Varga et al, 2005). However, not all the elements in the workflow can be processed automatically. For instance, generation of domain specific multilingual seed URL lists requires human involvement as well as the quality evaluation of the outputs from the Hunalign sentence aligner needs to be performed manually.

## 10. The LTO Catalogue

The LTO catalogue is not a new LR repository but a compilation of freely and easily accessible LRs that can be used for professional/commercial purpose. In addition, some LR that are not available for commercial purposes have been included when they were deemed of very high quality. LRs were selected according to the above mentioned usability criteria and, as far as possible, additional metadata was included if it was not easily available at the original repository. Furthermore, LRs were (and are continuously) checked by practitioners and commented on. MT developers found the comments function of the LR Catalogue more useful even than a rating system that will always remain arbitrary.

The LR are listed in the Catalogue in alphabetical order (based on their name as it appears in the original repository). Keywords and tags facilitate their searchability. The following metadata fields are available for each LR:

- Source name
- Author
- Resource name
- Description
- Languages (at present, a list of languages – it is foreseen to cater for the inclusion of language pairs),
- Resource link (in the original repository)
- Contact person
- Resource type
- Resource availability (for commercial purposes / free)
- Availability of direct download
- Modality (text, speech, spoken streams in video)
- Domain
- Technical format (with explanation)
- Size,
- Production date

Depending on the information made available in the original repository, not all these fields may have been completed. LR providers will be informed that metadata covering all these fields should be provided for LR to be included in the LT Observatory Catalogue.

## 11. Streamlining European Knowledge and Management of Language Resources

Commercial uses of parallel corpora language resources appear to be focused on individual end user cases or on the re-use of existing translation memories (owned by suppliers or organisations) if they are in-domain. There is obviously no exhaustive list of these resources available.

*Terminology* resources present a different case. There are a few large banks of terminology that are available for online use. Their main drawback is quality: users say there is out-of-date or erroneous content mixed in with high quality content. Terminology is now often shared by translators in repositories such as Proz, or can be searched online in the TAUS Data Cloud. But here again, there are no metadata that can provide quality controls on accuracy. IATE data can be downloaded via an API for MT use, but most term data cannot.

More generally, there are currently a number of different projects or efforts dedicated to identifying and networking language resources as a necessary asset for the multilingual digital single market. It is imperative that **there is overall agreement on the key goals of these projects and on the manner of evaluating success in reaching their various goals**. By working closely with both the translation/language industry (and its clients) and the academic and research community, LT-Observatory feels well-positioned to aid all parties concerned in reaching agreement on LR usability criteria which we believe will be most useful for starting the next stage of repository analysis, data collection, and inventing the next model of language data dissemination (sharing, market, crowd-sourcing/evaluating, etc.).

In this context, it also appears to be necessary to examine – yet again – the importance of ***legal constraints on data sharing (copyright, IPR, ownership, etc.)*** as there is still a lack of clarity in the minds of many about the rights of Europeans with respect to using LRs.

## 12. IPR issues

Automatic Web crawling for LT purposes involves the copying of content[2] and cannot be expected to fall under copyright exceptions and limitations in most laws in the EU[3]. Several investigations made on how to go about using web data lawfully have not given any answers or provided shortcuts. Web crawling will thus require separate permission from the rights holder. A case study conducted in the PANACEA[4] project, described in (Arranz & Hamon, 2012) sheds some light on the challenges and obstacles met in connection with obtaining permission to use Web data.

[2] "Data Protection Directive" (95/46/EC)

[3] See Article 5 of the EU Copyright Directive about exceptions and limitations.

[4] http://www.panacea-lr.eu/, see References

Web data is potentially very costly, depending on the number of sources that must be approached. Having managed to identify a contact point, which in itself can be quite difficult, negotiations about usage conditions must be carried out. In this process, the data owner will typically want to know what their content will be used for and they will often only accept a usage identical to their own intended usage. It is not unusual that many data owners are not familiar with concepts such as Human Language Technologies, machine learning etc. Given the complexity of the process, such negotiations can therefore be both lengthy and complicated.

**Discussion:** As essential components of languages, LRs should not be copyright subject matter - the raw material used to create LRs should be exempted from copyright protection for the narrow purposes of creating LRs and/or making languages interoperable.

The above principles should be established once and for all, particularly for LRs created on the basis of the reuse of public data and/or funded by public money.

There is a large consensus for the public sector to open up their databases for the creation of LRs. The EU should make proper curation and sustainable open access a requirement for data resulting from (publicly) funded projects. The copyright legislation should be clarified in such a way that the re-purposing of (textual) data where this does not affect the legitimate interests of the copyright owners is explicitly permitted. A revised European copyright law may, for example, accommodate an exception for the “decompilation of languages” for the purposes of developing language resources for machine translation, i.e. a provision transposed *mutatis mutandis* from the "reverse engineering/decompilation" exception presently available in the EC Software Directive - art. 6).

## 13. Results

The most important conclusion is that many resources exist which are useful for MT and similar work, but the majority are for (academic) research or educational use only, and as such **not available for commercial use**.

If companies have collected useful language resources for their own purposes, this is an asset that they do not easily share with their competitors. During the user study companies expressed that if they need LR they search the web and use what they find, this is often the fastest and cheapest way. It should be noted that it may happen that some of the resources collected this way are actually not available for commercial or any other use because of IPR problems.

Around 100 relevant resources were collected in the first year. The identified resources can be divided into parallel corpora (36), comparable corpora (9), monolingual corpora (10), thesauri (3), speech corpora (4), glossaries (8), terminological resources (22), tools (4), lexicon (1) and treebanks (4).

### Corpora

Parallel corpora are here corpora where the same text appears in more than one language. In Figure 1 we give for each language the number of corpora in which it appears.

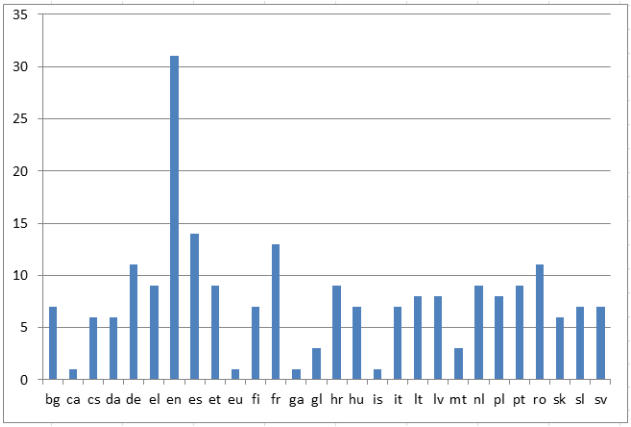


Figure 1: Number of parallel corpora per language

The most frequent language by far is English as it appears in 31 parallel corpora out of 36 in all. Other frequent languages are Spanish, French, German and Romanian appearing in 14, 13, 11 and 11 corpora respectively. These languages in different combinations also constitute the most frequent language pairs. It should however be noted that the corpora where these languages appear together are mostly multilingual - and the languages are therefore not language pairs in the sense where a source and a target language can be identified.

Other languages with a medium frequency are Greek, Estonian, Portuguese and Croatian as they appear in 9 corpora each.

Out of the 36 parallel corpora 14 are bilingual (the rest are multilingual) and these all have English as one of the languages, except one where the language combination is French-Dutch. In the remaining 13 resources Croatian, Portuguese, Greek, Hungarian and Latvian are among the languages that constitute either the source or target language together with English.

The project also collected 9 comparable corpora mostly comprising East European languages and English.

The corpora comprise 19 subject domains of very different character, broad domains e.g. health/medicine and narrow e.g. building foam and sealant. Since the metadata of the resources are not based on a classification scheme, some standardization is needed. We made a comparison with EUROVOC[5] which shows that some resource domains – e.g. administration – are spread over several EUROVOC subdomains og micro thesauri whereas others such as politics are equivalent to a single EUROVOC top domain. Not all of the EUROVOC top domains are represented in the resource metadata but since some of the resources covering many domains do not specify all the included domains, the domain coverage

[5] http://eurovoc.europa.eu/

is broader than what appears from the metadata.

## Terminology

The terminological resources and thesauri collected in the scope of LTO are bilingual and multilingual. There are no monolingual resources. Only 13 percent of the resources are bilingual (as depicted in Figure 10), all but one of them covering the language combination French-English. The majority of the LTO resources are multilingual (88 percent).

English is the only language represented in all of the terminological resources, closely followed by French (91 percent of the terminological resources). German, Spanish, Italian, Portuguese, and Swedish are represented in more than 10 terminological resources. Greek, Finnish, and Polish are covered in 10 resources. Bulgarian, Czech, Danish, Estonian, Irish, Croatian, Hungarian, Latvian, Lithuanian, Maltese, Dutch, Romanian, Slovak and Slovenian are represented in less than 10 terminological resources in the LTO collection (see Figure 2).

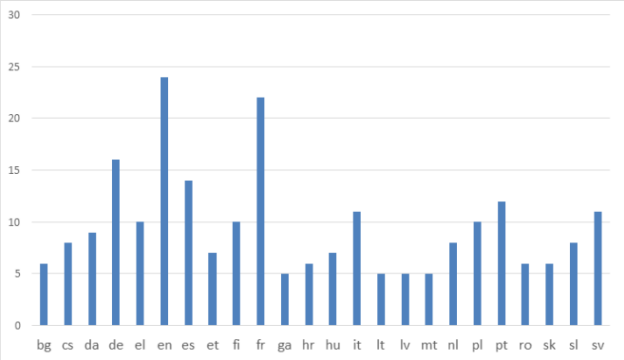


Figure 2: Terminology LR Distribution per language

## 14. Gap description

The work done in the Language Technology Observatory project has revealed a number of gaps. Below we suggest considering gaps along several dimensions:

**Coverage gap**

*Corpora:* Only English has a good coverage in terms of combinations with other languages. Spanish, German, French, Latvian, Romanian, Croatian, Polish and Lithuanian are moderately covered in relation to other languages. Maltese, Danish, Czech and Slovak are poorly covered. All languages, including English have gaps in relation to domains. EUROVOC top categories not mentioned in the resource metadata for any language are: Trade, Finance, Transport, Agriculture Forestry and Fisheries, Production, Technology and Research, Geography, International Organizations. Besides, many subdomains within many other top categories are not represented in any or only in very few languages.

If we compare the META White Papers' support level cf. http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison with the support level we have arrived at in the LTO project, we see that even if the overall picture is similar - English is still reasonably well served as the only language - there are also some developments: The moderately supported group is different today as Latvian, Romanian, Estonian and Lithuanian are included whereas Czech and Swedish are in the fragmentarily supported group.

*Terminology:* In terms of the number of resources in which a language is represented, English, followed by French, is the most represented in terminology resources, as both languages are often used as pivot languages. German, Spanish, Italian, Portuguese, and Swedish are moderately represented in terms of the number of resources in which they can be found. Greek, Finnish, and Polish are less represented in terminological resources. Bulgarian, Czech, Danish, Estonian, Irish, Croatian, Hungarian, Latvian, Lithuanian, Maltese, Dutch, Romanian, Slovak and Slovenian are the least represented languages in the terminological resources.

Nevertheless, the number of the resources covering a language should not be the sole criterion for coverage. The size of the resource in relation to the domains covered should also be considered. In this regards, smaller languages with national terminology infrastructures or terminology centres (e.g. Swedish, Catalan, Irish) have better coverage than only the number of available resources would suggest.

**Awareness gap**

There is clear evidence that a large part of the demand side is unaware of the offer. When the project made the first 30 pre-selected LRs available, it appeared that a large segment of potential users was not aware of their existence. Our initiative was therefore welcomed by these actors. This was also amply confirmed by the very interested and positive reception of Ralf Steinberger's presentation of the JRC's LRs at the LT-Accelerate conference on 23-24 November 2015. Clearly, there is a need to reach out to the demand side and to "market" the offer to it in a more proactive fashion.

**Quality gap**

Already at pre-selection stage, it became clear that very few LRs that are presently on offer in existing repositories correspond to minimum quality requirements on metadata. Moreover, it appeared that only VERY FEW LRs are made available by repositories in a way that enables their straightforward commercial use. The latter is often restricted in the first place; licensing conditions are not clearly spelled out; contact persons to obtain additional information are not identified; where LRs are made available for a price, the quality-price relationship is often considered inadequate; etc. Hence, there is a need to improve the quality of existing LRs and to reflect seriously on the conditions at which they can/should be made available for commercial use (particularly when they have been compiled with the support of public money).

**Quantity gap**

For LRs to be useful in an operational context they need to be available in large quantities. It is obvious that the quantities available today (at the required quality levels) are largely insufficient to have a positive impact on the quality of MT in a commercial context. A large combined effort should be launched to produce new LRs across the board (and in all languages) that correspond to a set of

agreed quality criteria. This would also (and URGENTLY) require a clarification of their copyright status in a commercial context. Furthermore, there is demand for in-domain resources, i.e. LRs that are clearly customised for use in specific domains (healthcare, finance, security, tourism, etc.). Whether the compilation of such in-domain resources should be left to private initiative as is currently the case or whether they should become part of a European Language Cloud should be, at the very least, debated seriously. There is evidence that only very few European commercial players will be able to compile such in-domain resources in sufficient quantities and qualities over the long-run to allow them to offer specialised domain-specific language clouds. Furthermore, the question whether such specialised language clouds should be left to private appropriation in the first place should also be debated.

## 15. Future steps

For the future, a forward-looking strategy should be devised involving the following steps:

- Identify all relevant resources that satisfy "operational usability" defined above and promote them through the LR Catalogue,
- Create synergy between all projects touching upon LRs,
- Encourage a serious effort to make LRs commercially available (particularly when they have been created with public funding) on the basis of credible and sustainable business models
- Support an effort by the EU to clarify once and for all the legal situation of LRs.

As we have seen above, only one language (English) has a reasonable coverage in relation to volume as well as in relation to domains, and a very limited number of languages have a moderate support. The reason for this may be that EU language resource identification and management has been a pretty random process, unsupervised for several years (decades) despite best intentions.

From now on, LR identification and operational management needs to be organized by means of a clear **strategy** of identifying, quality-checking and promoting all those LRs that can contribute to better MT productivity in the years ahead[6]. The technology that can help enable this provision of LRs is itself developing by automatic methods of creating parallel corpora e.g. from crawling the web, cf. section 9 on Tools.

New methods of categorizing the meaning of words and sentences across multiple languages are opening up new opportunities for more effective resources for MT, so a bit further into the future, we need to explore how these results can also be used for improving the quality and the quantity of LRs for MT. And if possible, clearly align a given tranche of technology R&I with LR usability needs.

[6] Later a plan will be needed for better productivity in other important areas, apart from MT.

It will take several cycles for MT selection, use, and refinement to make the most of what exists. Tools will need to be developed that can

- Help the quality-checking of existing LRs.
- Boost LR creation (automatic methods, semantic categorization etc.)
- Help identifying the usability (domain relevance and quality) of any given resource/language pair etc.

LT Observe is taking the first step: simplifying access to usable (and preferably free) LRs from public repositories in the EU via a one-stop access point. Once the LT Observatory catalogue is open for business, there should be a pilot study of usability, with feedback from users to improve the service for a second round of LR collection/invitations/pooling/crowdsourcing opinion.

## 16. Acknowledgements

The LT_Observatory project is supported by the European Commission under grant agreement 644583. We want to thank all participants in the project for their support, and in particular Hanne Fersøe, University of Copenhagen, for input on IPR.

## 17. References

Arranz V, & Hamon O. (2012). On the Way to a Legal Sharing of Web Applications in NLP. *Proceedings of LREC 2012*, Istanbul, Turkey.

Fersøe, H., Monachini, M.(2004): ELRA Validation Methodology and Standard Promotion for Linguistic Resources. *Proceedings of the fourth International Conference on Language Resources and Evaluation (LREC2004),* ELRA, Lisboa.

Mastropavlos N., & Papavassiliou V. (2011). Domain Adaptation of Statistical Machine Translation using Web-Crawled Resources: A case study. *Proceedings from the 10th International Conference of Greek Linguistics*.

META-NET White Paper Series (2012): *Europe's Languages in the Digital Age*. 32 Volumes. Eds. Rehm, Georg, Uszkoreit, Hans, Springer.

Pecina P., Toral T., Papavassiliou V., Prokopidis P., & Genabith van J. (2012). Domain Adaptation of Statistical Machine Translation using Web-Crawled Resources: A case study. *Proceedings from the 16th EAMT Conference.*

Varga D., Németh L., Halácsy P., Kornai A., Trón V., & Nagy V. (2005). Parallel corpora for medium density languages. *Proceedings of the RANLP*.

List of the catalogues and projects used or consulted:
http://clarin.eu/content/virtual-language-observatory),
http://www.meta-net.eu/meta-share
http://www.elra.info/en/catalogues/
http://opus.lingfil.uu.se/
https://www.taus.net/
https://www.letsmt.eu/Systems.aspx
http://www.lider-project.eu/
http://falcon-project.eu/
http://www.panacea-lr.eu/,
http://www.ttc-project.eu/
http://cesar.nytud.hu/about/

www.eurotermbank.com
https://ec.europa.eu/jrc/
https://ec.europa.eu/jrc/en/language-technologies/.
http://iate.europa.eu/
http://eurovoc.europa.eu/
http://dublincore.org/documents/1999/07/02/dces/
http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus
http://www.tnc.se/
http://www.tearma.ie/
http://www.termcat.cat
http://www.uzei.eus/
http://struna.ihjj.hr/
http://www.evroterm.gov.si/
http://www.tsk.fi/tepa
http://www.data.gouv.fr/fr/datasets/base-franceterme-termes-scientiques-et-techniques/
http://www2.cfwb.be/franca/xml/html/bd/bd.htm
http://www.eionet.europa.eu/gemet
http://glossary.eea.europa.eu/terminology/
http://ec.europa.eu/dgs/home-affairs/what-we-do/networks/european_migration_network/docs/emn-glossary-en-version.pdf
http://www.fao.org/biotech/biotech-glossary/en/
http://untermportal.un.org/